An analysis of student-teacher information propagation

Jeremy H. M. Wong

Department of Engineering, University of Cambridge
Trumpington Street, CB2 1PZ Cambridge, England

jhmw2@cam.ac.uk

Abstract

In the field of Automatic Speech Recognition (ASR), ensemble methods have often been found to give significant performance improvements over single models [1]. Ensemble methods can be viewed as a Monte Carlo approximation to Bayesian inference, by taking a combination over a finite number of models. When constructing an ensemble, it is important to consider the forms of diversity used, as this influences the ability of the ensemble to capture uncertainty about the model. Some examples of methods for generating diverse models include using random weight initialisations [2], Dropout [3], random forests [4], and different model types [5]. It is also important to consider the computational cost of performing recognition through the ensemble. This tends to scale with the ensemble size, and can be an issue when trying to implement models on devices with limited hardware. As a whole, this PhD project is concerned with investigating methods to generate diverse models within the ensemble, as well as methods to improve the computational efficiency of performing recognition using the ensemble.

One possible method to reduce this computational cost is student-teacher training [6]. A single student model can be trained to emulate the combined behaviour of the ensemble. During the recognition stage, only this single student model needs to be used. The student can be trained to emulate the combined ensemble posteriors at either the frame [7] or hypothesis [2] level. At the frame level, a standard method of propagating information from the ensemble of teachers to the student is to minimise the KL-divergence between their frame posteriors [7]. It has often been found that a student trained with sufficiently powerful teachers is able to perform better than one trained on forced alignment hard targets. It is an interesting question to ask, what is it about the teachers' posteriors that is beneficial for the student, in addition to the information that is already available in the hard targets. This paper is focused on a part of the PhD project that aims to address this question.

One hypothesis is that the teachers' posteriors include information about how confusable the teachers believe that a frame is. This can manifest, for example, as higher entropy posteriors or by more teachers disagreeing on the classification of a frame. With this information, the student can be trained with less emphasis placed on correctly classifying frames that are inherently confusable. This may allow the student to be trained more easily. It is interesting to analyse how much the student can gain from the different types of frames.

The work in [8] performed such an analysis when having a single teacher, by categorising the training data frames into those that the teacher correctly and incorrectly classifies. The student, trained on each of these categories, was able to gain most of its performance from the frames that were misclassified by the teacher. This paper extends this analysis to an ensemble of teachers. With an ensemble, it is possible to subdivide the frames more finely into multiple categories with different levels of confusability. The frames can be categorised, depending on how many teachers classify them correctly and how many teachers agree with each others' classifications. The contributions of each of these categories of frames to the student performance can be accessed by varying for each category, whether the forced alignments or teachers' posteriors are used as targets, or whether the category is included in training.

This analysis is particularly interesting when there are architectural differences between the teachers and the student. Two architectural differences investigated in this paper are where the teachers and student differ in either their model type or output targets. In a hybrid ASR architecture, there are many possible choices for the type of neural network for the acoustic model. This paper investigates propagating information between feed-forward deep neural network and long short-term memory models. Also in a hybrid ASR architecture, the outputs of the neural networks usually represent clusters of context dependent phone states. The contexts need to be clustered together in order to reduce the number of parameters, to allow for more robust models. The clustering process can be modified to generate a variety of sets of state clusters, by for example using the random forest method [4]. An ensemble can be constructed by associating a separate neural network with each set of state clusters [9]. The set of state clusters used for the student model can be chosen independently of those of the teachers. In these situations, it is then useful to determine what behaviours of the teachers the student can effectively capture.

Index Terms: Ensemble, student-teacher, speech recognition

Acknowledgements

This work has been performed under the generous supervision of Professor Mark Gales, at the University of Cambridge.

1. References

[1] J. G. Fiscus, "A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER)," in *ASRU*, Santa Barbara, USA, Dec 1997, pp. 347–354.

[2] J. H. M. Wong and M. J. F. Gales, "Sequence student-teacher training of deep neural networks," in *INTERSPEECH*, San Francisco, USA, Sep 2016, pp. 2761–2765.

[3] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, pp. 1929–1958, Jun 2014.

[4] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, Cagliari, Italy, Jun 2000, pp. 1–15.

[5] L. Deng and J. C. Platt, "Ensemble deep learning for speech recognition," in *INTERSPEECH*, Singapore, Sep 2014, pp. 1915–1919.

[6] C. Bucilă, R. Caruana, and A. Niculescu-Mizil, "Model compression," in *KDD*, Philadelphia, USA, Aug 2006, pp. 535–541.

[7] J. Li, R. Zhao, J.-T. Huang, and Y. Gong, “Learning small-size DNN with output-distribution-based criteria,” in *INTERSPEECH*, Singapore, Sep 2014, pp. 1910–1914.

[8] J. Cui, B. Kingsbury, B. Ramabhadran, G. Saon, T. Sercu, K. Audhkhasi, A. Sethy, M. Nussbaum-Thom, and A. Rosenberg, “Knowledge distillation across ensembles of multilingual models for low-resource languages,” in *ICASSP*, New Orleans, USA, Mar 2017, pp. 4825–4829.

[9] T. Zhao, Y. Zhao, and X. Chen, “Building an ensemble of CD-DNN-HMM acoustic model using random forests of phonetic decision trees,” in *ISCSLP*, Singapore, Sep 2014, pp. 98–102.